上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Mixup Training for Generative Models to Defend Membership Inference Attacks

Zhe Ji[1], Qiansiqi Hu[1], Liyao Xiang[1], Chenghu Zhou[2]

[1]*Shanghai Jiao Tong University*
[2]*Institute of Geographic* Sciences and Natural Resources Research, CAS

饮 水 思 源 · 爱 国 荣 校

Speaker: Zhe Ji

Content

01-Background

Scenarios, MIA threats, existing works.

02- Preliminaries

The likelihood ratio attack, mixup training.

03-Methodology

Defense algorithm, analytical insights.

04-Experiments

Privacy results and utility results.

01

Background

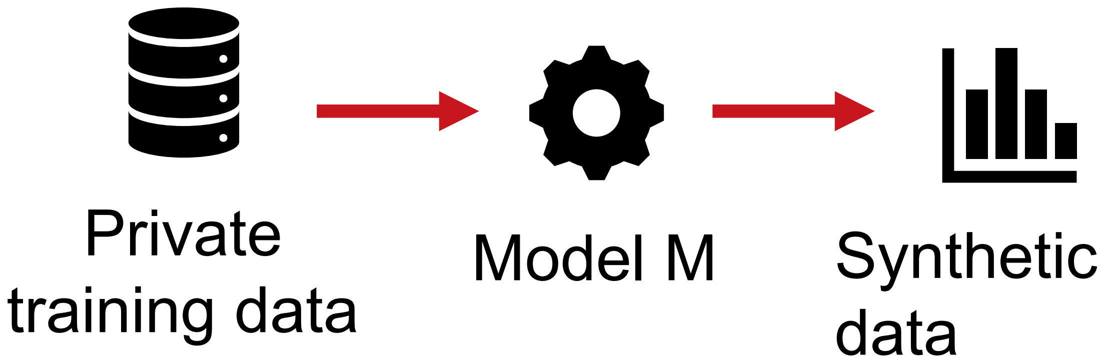
MIA threats, scenarios, existing works.

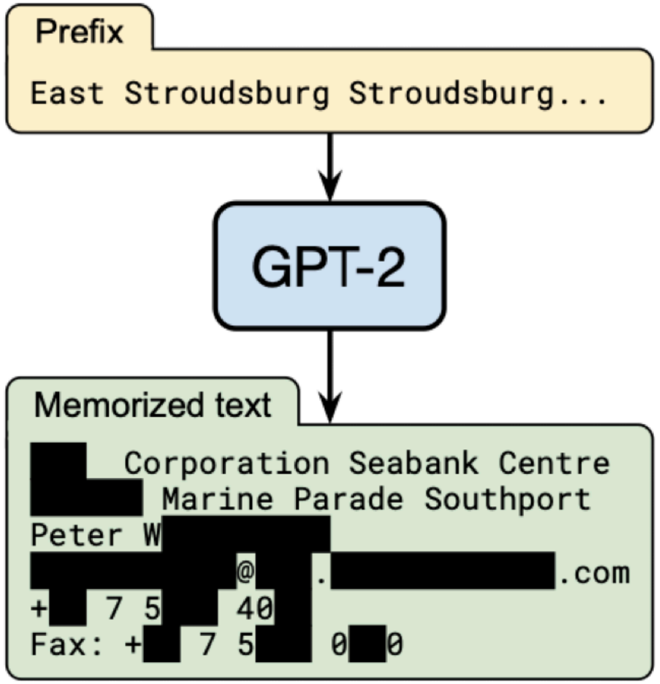
Setup

Training dataset: **Private** data

Object to release: Trained model, or synthetic data from the trained model

Threats: The model may **memorize** training samples. Then attackers may **recover**, or **infer** the private training data from the released model or synthetic data





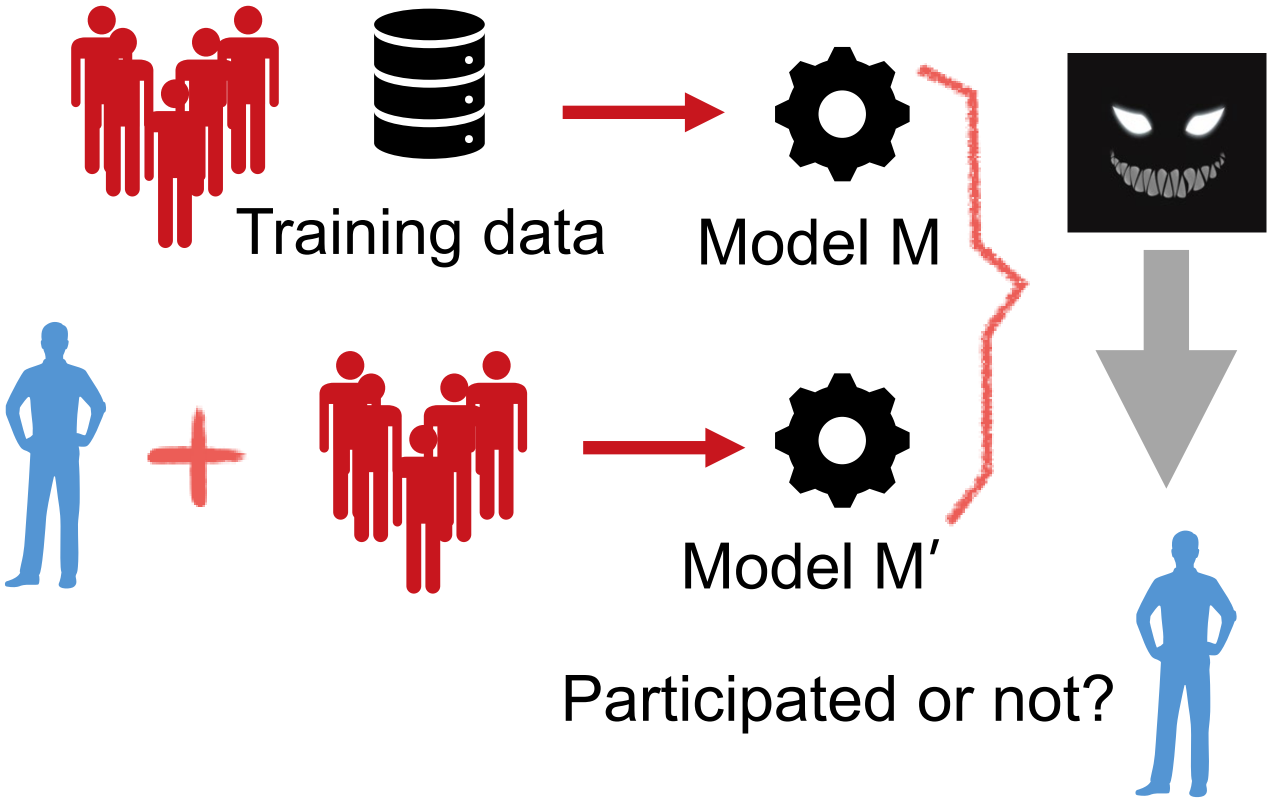
Membership Inference Attack (MIA)

Attacker input:

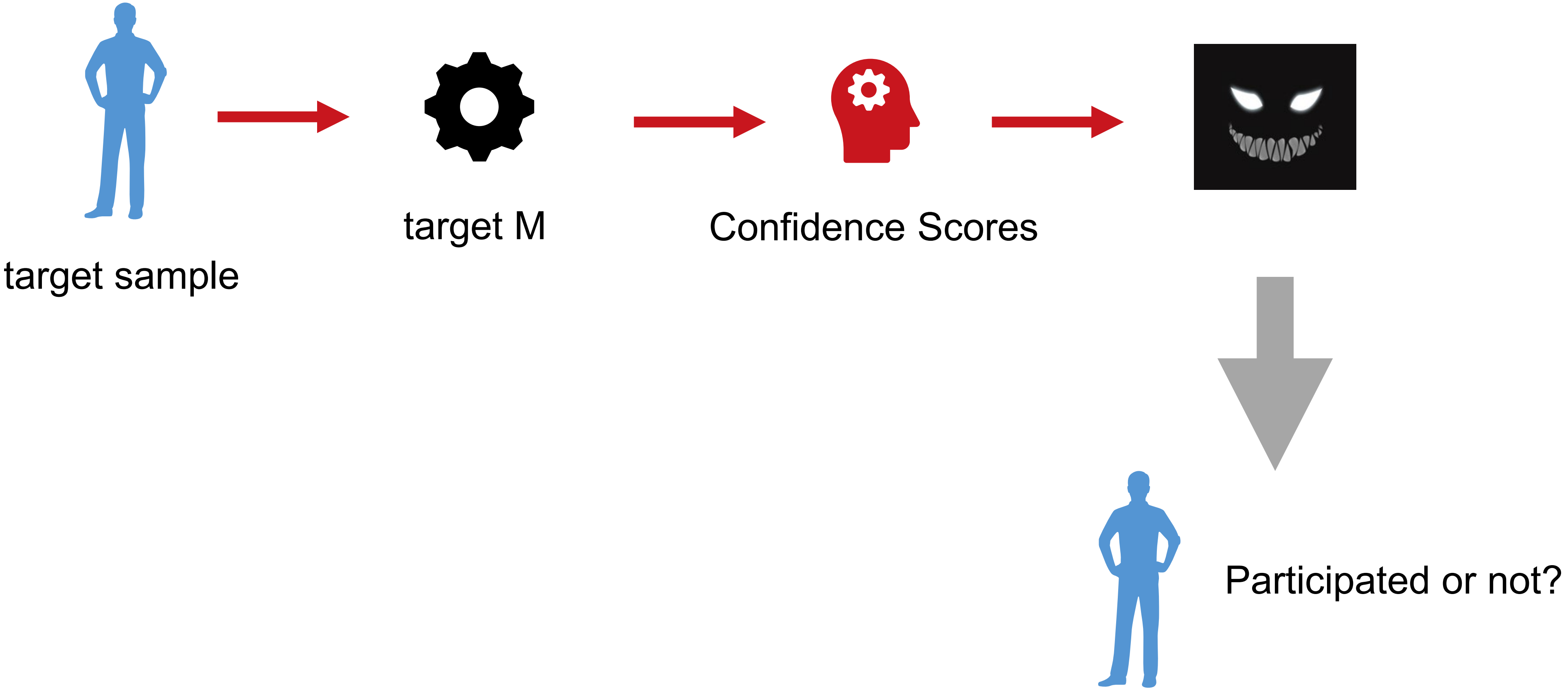
- A target model (victim)
- A target sample

Attacker output:

- The predicted probability that the target sample belongs to the training dataset of the target model

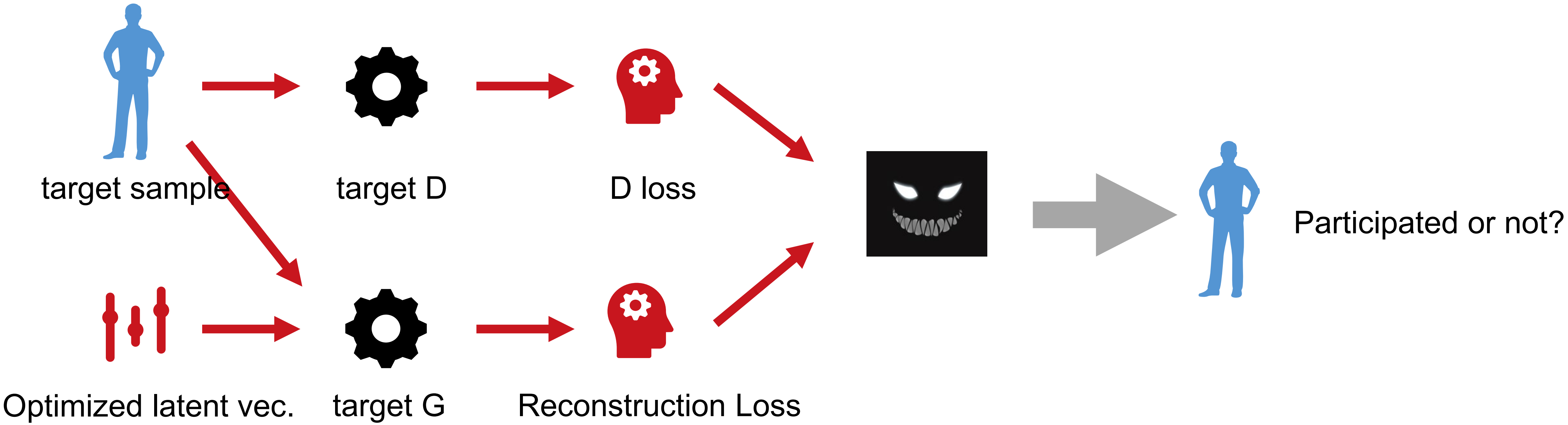


MIAs against Discriminative Models



MIAs against Generative Models

Assuming target M is a GAN, containing a generator G and a discriminator D



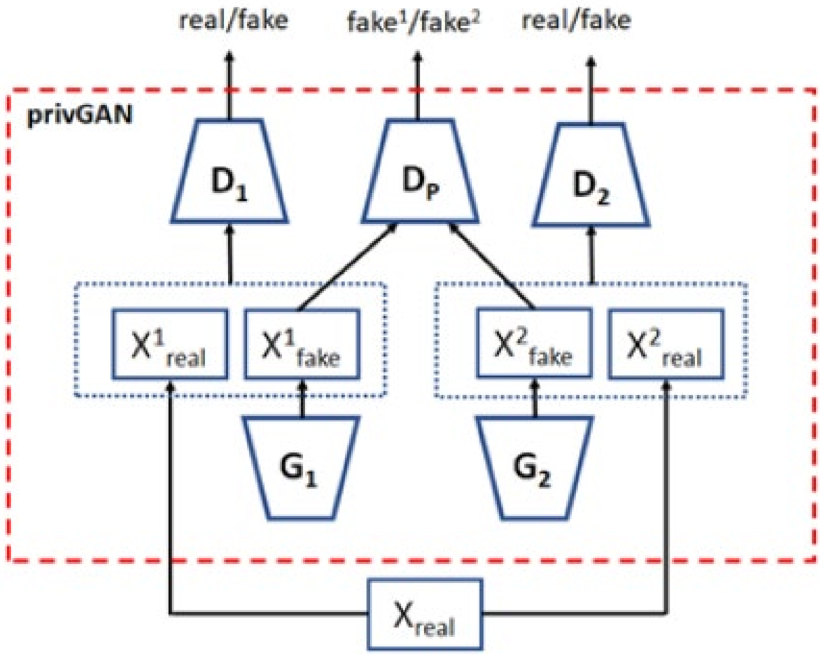
Defense against MIA on Generative Model is Hard

- **Features differed from discriminative models:**
  1. No confidence scores as ouput
  2. Unknown downstream tasks

- **Goals of defense:**
  1. No reproduction or memorization of training data
  2. Data utility reservation

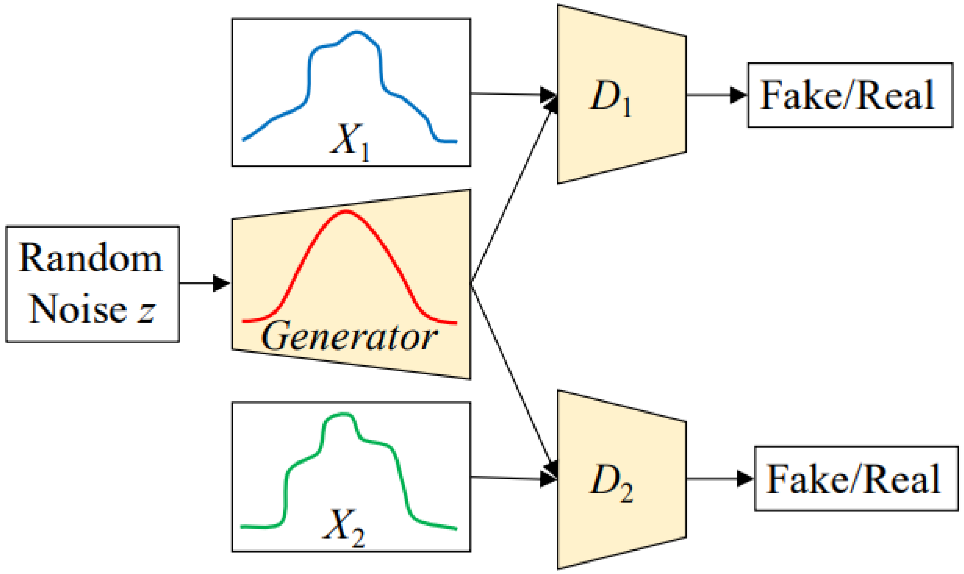
Existing Solutions

PrivGAN



D_p = built-in adversary to predict which generator produces a synthetic sample

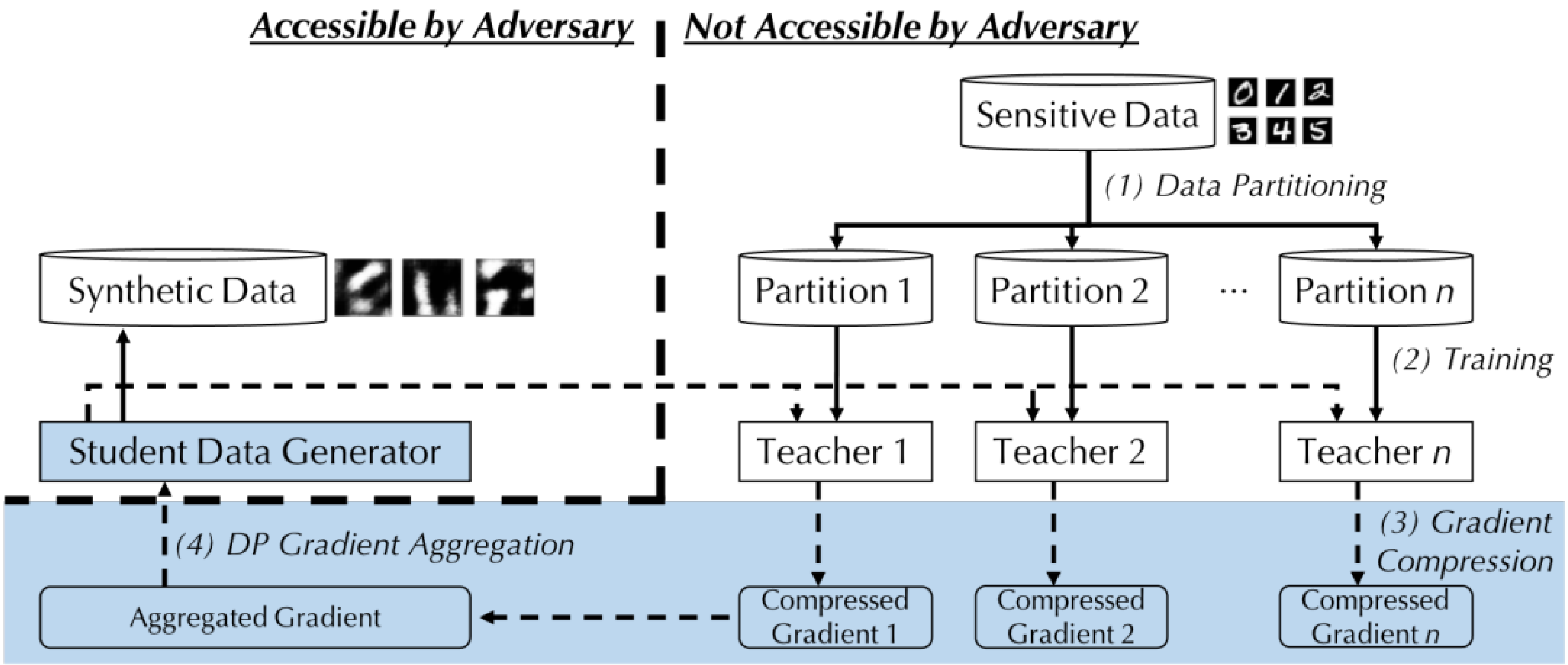
PAR-GAN



J. Chen, W. H. Wang, H. Gao, and X. Shi, "Par-gan: Improving the generalization of generative adversarial networks against membership inference attacks," in Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 127–137.
S. Mukherjee, Y. Xu, A. Trivedi, and J. L. Ferres, "privgan: Protecting gans from membership inference attacks at low cost to utility," Proceedings on Privacy Enhancing Technologies, vol. 2021, no. 3, pp. 142–163, 2021. [Online]. Available: https://doi.org/10.2478/ popets-2021-0041

Existing Solutions

DataLens



TopAgg: noisy gradient compression and aggregation

B. Wang, F. Wu, Y. Long, L. Rimanic, C. Zhang, and B. Li, “Datalens: Scalable privacy preserving training via gradient compression and aggregation,” in Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, 2021, pp. 2146–2168.

Problems of Existing Solutions

- **Some works focus on the design of GAN architectures**
  - E.g. PAR-GAN, privGAN
  - Shortcoming: complexity, computation overhead
- **Some use differential privacy**
  - E.g. Datalens
  - Shortcoming: utility degradation

None has considered the strongest MIA, LIRA.

02

Preliminaries

The likelihood ratio attack, mixup training.

The Likelihood Ratio Attack (LIRA)

Likelihood ratio

H0: the target example is a member.

H1: the target example is not a member.

$$\Lambda(M) := \frac{\Pr(M|M_{in})}{\Pr(M|M_{out})}$$

Implementation

LIRA replaces distributions of models with distributions of losses, denoted by Q_{in} and Q_{out}

$$\Lambda(l) := \frac{\Pr(l|Q_{in})}{\Pr(l|Q_{out})}$$

LIRA focuses on the true positive rate (TPR) at low false positive rate (FPR) regime

The Likelihood Ratio Attack (LIRA)

Λ
9.4
5.2
4.4
3.5
1.0
0.8
0.7
0.6
0.5

Threshold

Red: Member
Black: Non-member

After attacking several target samples, attackers can choose a relatively high threshold to reach a low FPR

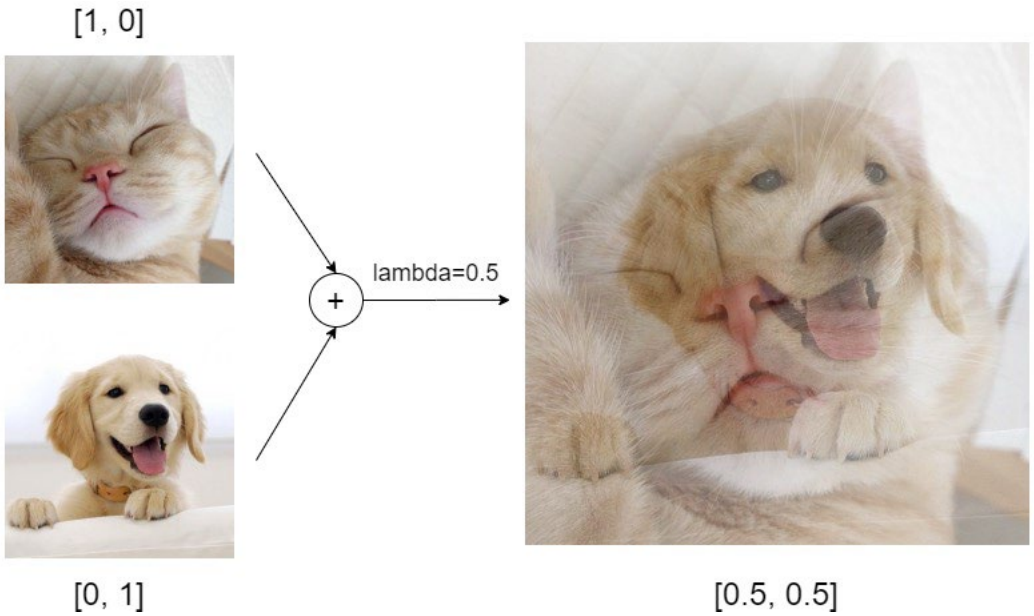
Here TPR=3/5, FPR = 1/4

Existing defenses all fail to reduce TPR at low FPR

We target at this threat

Mixup Training

Mixup training regularizes the neural network to favor simple linear behavior in between training examples



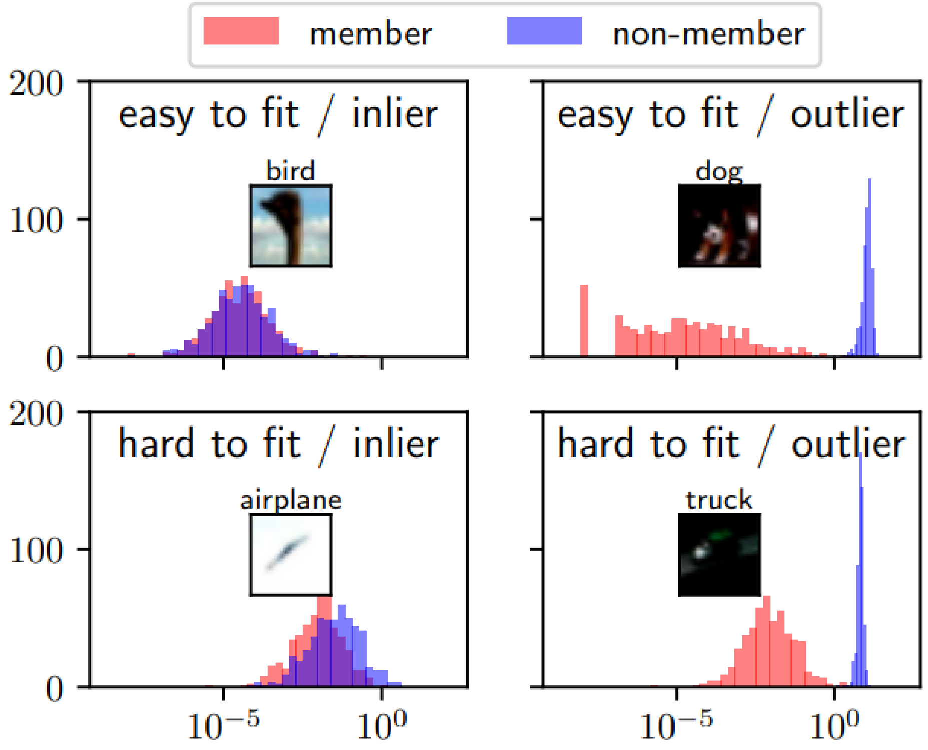
Original data → Mixup data → Normal training

03

Methodology

Defense algorithm, analytical insights.

Intuition



Outliers:

1. Significant influence to the target model
2. Easy to be detected by MIA attackers

$$\Lambda(l) := \frac{\Pr(l|Q_{in})}{\Pr(l|Q_{out})}$$

Intuition:
Reduce the influence of outliers

The figure above comes from the paper: N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, “Membership inference attacks from first principles,” in 2022 IEEE Symposium on Security and Privacy (SP). IEEE Computer Society, 2022, pp. 1519–1519.

Mixup Training

We use mixup to reduce the impact of outliers, so that the model does not differ greatly (in terms of loss) between members and non-members

randomly sample $(x_1, y_1), (x_2, y_2)$ from $\mathcal{D}$;
sample $\lambda \sim \beta(\alpha, \alpha)$;
$x_{mix} = \lambda x_1 + (1 - \lambda) x_2$;
$y_1 = one\_hot(y_1)$;
$y_2 = one\_hot(y_2)$;
$y_{mix} = \lambda y_1 + (1 - \lambda) y_2$;
/* generate fake samples */
sample $z \sim P_z$;
$fake = G(z, y_{mix})$
/* update D */
$L_D = -D(x_{mix}, y_{mix}) + D(fake, y_{mix})$;
$\theta_D = \theta_D - lr_D \cdot \nabla_{\theta_D} L_D$;
$batch\_done = batch\_done + 1$;
/* update G */
if $batch\_done \mod n_g == 0$ **then**
  $L_G = -D(fake, y_{mix})$;
  $\theta_G = \theta_G - lr_G \cdot \nabla_{\theta_G} L_G$;
end

Analytical Insights

PDF of loss value of members is shifted right by mixup

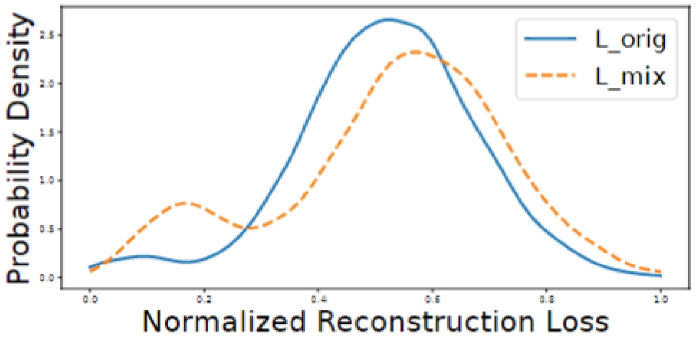


Fig. 1: Reconstruction Loss Distributions on CIFAR-10

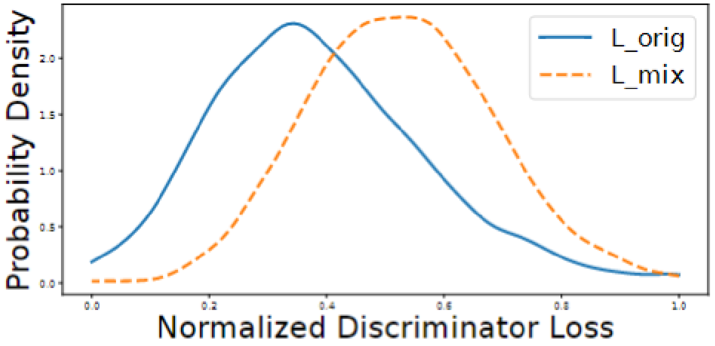
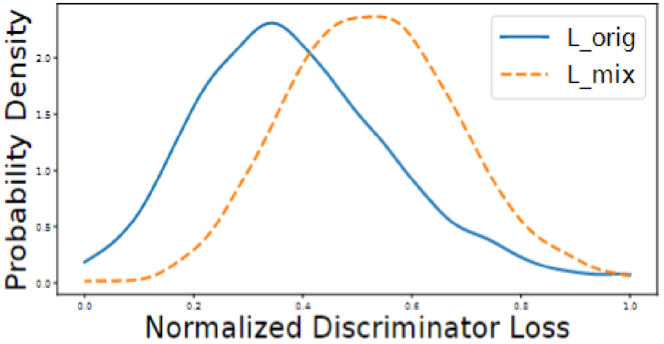


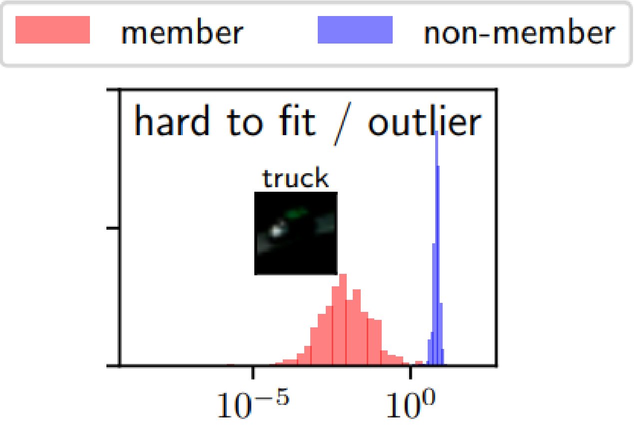
Fig. 2: Discriminator Loss Distributions on MIMIC-III

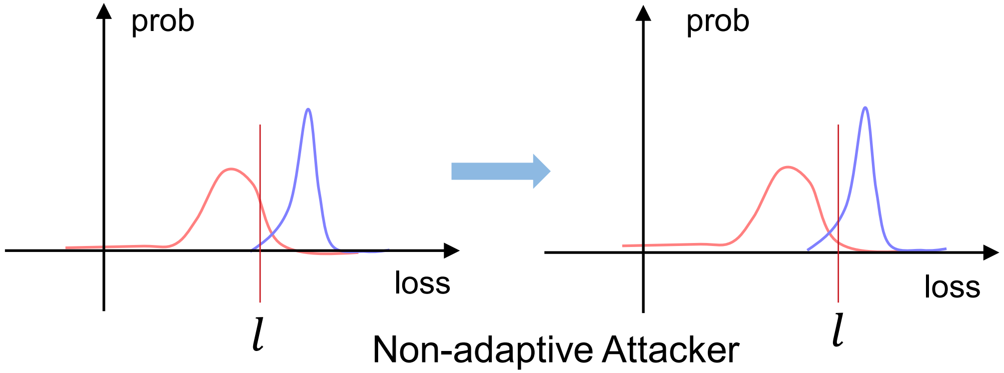
We will see how $\Lambda = \frac{\Pr(l|Q_{in})}{\Pr(l|Q_{out})}$ changes, intuitively and theoretically.

Analytical Insights

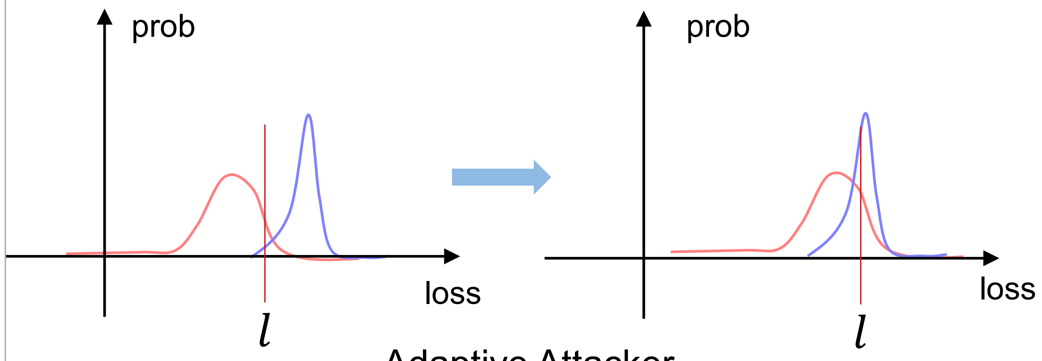


$$\Lambda = \frac{\Pr(l|Q_{in})}{\Pr(l|Q_{out})}$$





Non-adaptive Attacker



Adaptive Attacker

Analytical Insights

Ratio Λ is reduced for targeted members.

$$\Lambda = \frac{\Pr(l|Q_{in})}{\Pr(l|Q_{out})} \propto \exp\left(\frac{(l-\mu_{out})^2}{2\sigma_{out}^2} - \frac{(l-\mu_{in})^2}{2\sigma_{in}^2}\right)$$
$$\propto \exp\left[\left(\sigma_{in}^2 - \sigma_{out}^2\right)l^2 + 2\left(\mu_{in}\sigma_{out}^2 - \mu_{out}\sigma_{in}^2\right)l\right]$$
$$\stackrel{\text{def}}{=} \exp\left[f(l)\right]$$

Proposition 1:

With a probability approximately larger than 0.5, applying mixup to the model training leads to a decrease in Λ for target members.

Proof: by discussing the sign of $\sigma_{in}^2 - \sigma_{out}^2$.

Analytical Insights

Conclusion: Mixup training lowers the upper bound of attack AUC.

Symbols:

- P_m (or P_n): Distribution of Λ of members (or non-members)
- Q_m (or Q_n): Distribution of $\log \Lambda$ of members (or non-members)
- $\Xi = \log \Lambda$

$$AUC \leq -\frac{1}{2} D_{TV}(P_{\mathrm{m}}, P_n)^2 + D_{TV}(P_{\mathrm{m}}, P_n) + \frac{1}{2}$$

Lemma: Decreasing Λ for target members -> Upper bound of $D_{TV}(Q_m, Q_n)$ decreases.

Proof: Q_m, Q_n are Gaussians -> D_H, u. b. of D_{TV} , has CLOSED FORM about Λ

Analytical Insights

About Q_m , Q_n:

$\sigma_{in}^2 - \sigma_{out}^2 = 0$, when Q_m , Q_n are Gaussians.

Other cases:
Experiments show distribution of Ξ resembles a Gaussian (right figure)

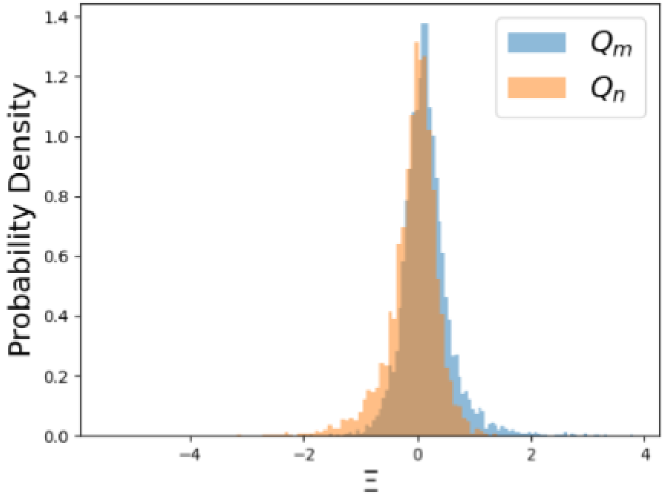


Fig. 3: Distribution of Ξ when $\sigma_{in}^2 - \sigma_{out}^2 \neq 0$.

04

Experiments

Privacy results and utility results.

Settings

Datasets

Images:
1. CelebA
2. CIFAR-10

Tables:
MIMIC-III

MIAs

GAN-leaks (against G)
Logan (against D)
LIRA (both)

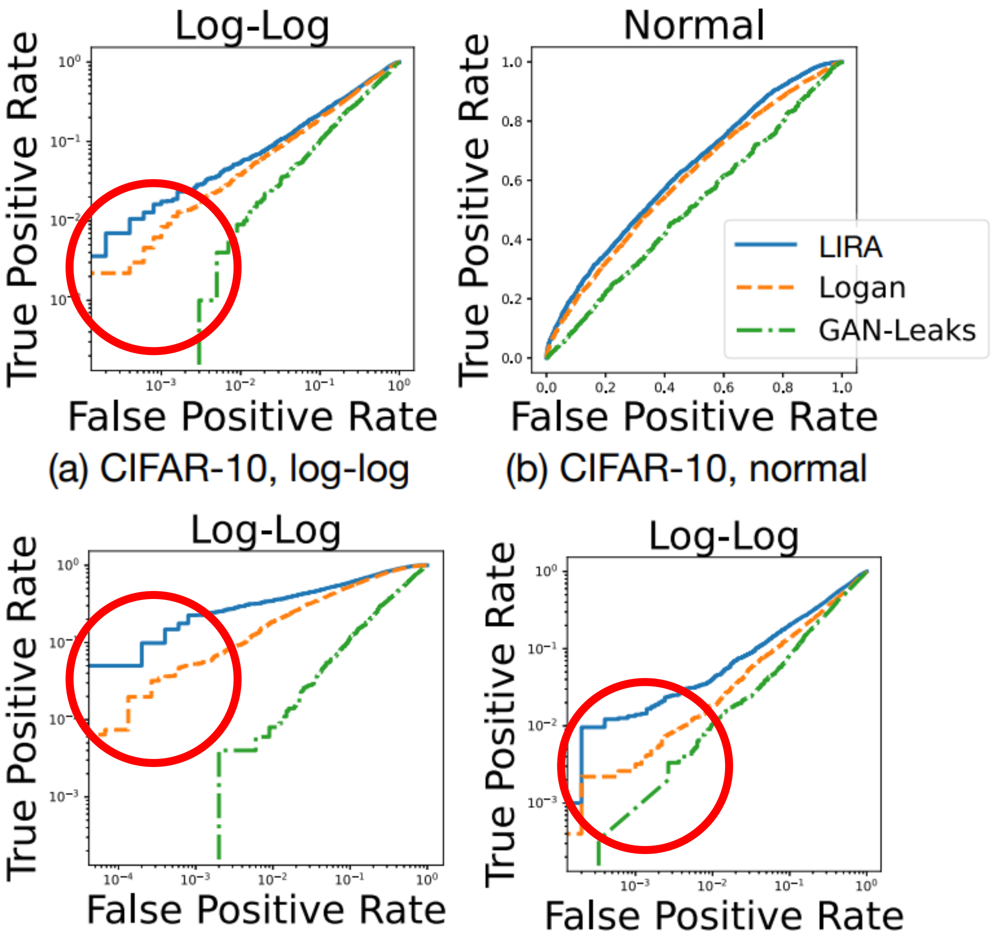
Defenses

Baselines:
1. PAR-GAN
2. RelaxLoss

Metrics

- Utility:
  - Downstream classification accuracy
  - Frechet Inception Distance for images
  - Dimensional Wise Probability for tables
- Privacy:
  - Area under ROC curve of MIA

Comparing Attacks



(a) CIFAR-10, log-log (b) CIFAR-10, normal

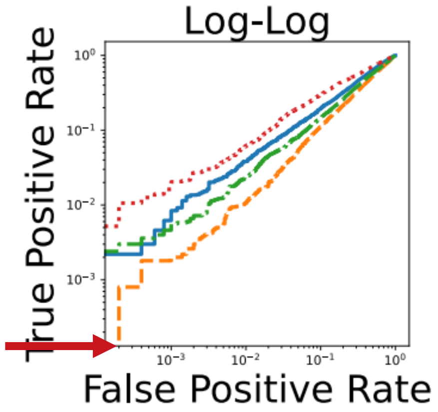
(c) CelebA (d) MIMIC-III

LIRA is the most powerful attack algorithm, from both perspectives:

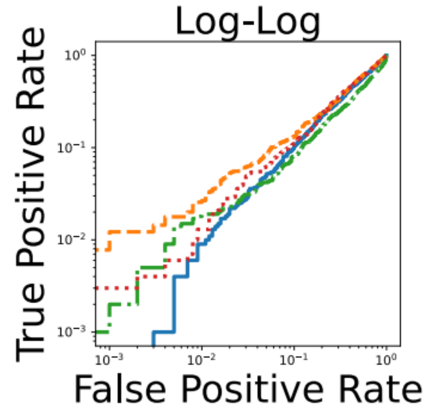
1. TPR when FPR is low

2. Area Under ROC Curve

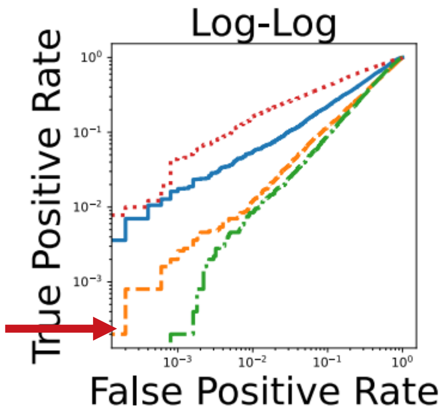
Privacy Performance: ROC Curve



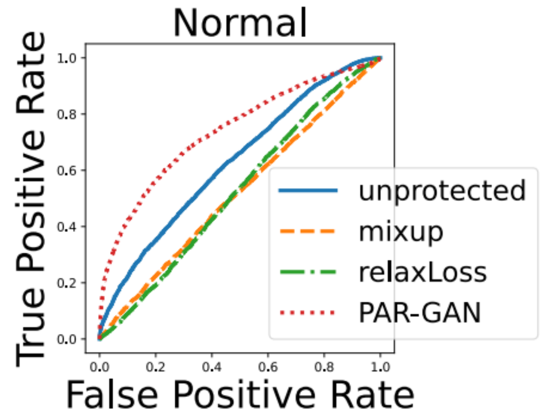
(a) Logan



(b) GAN-Leaks

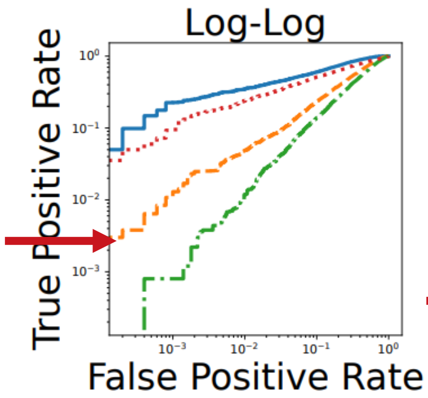


(c) LIRA, log-log

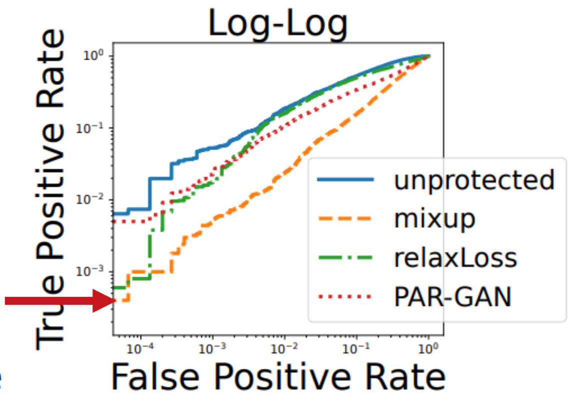


(d) LIRA, normal

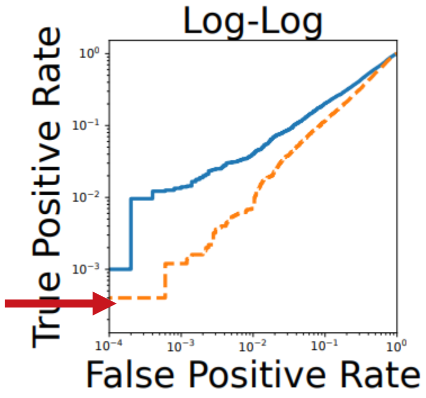
CIFAR-10



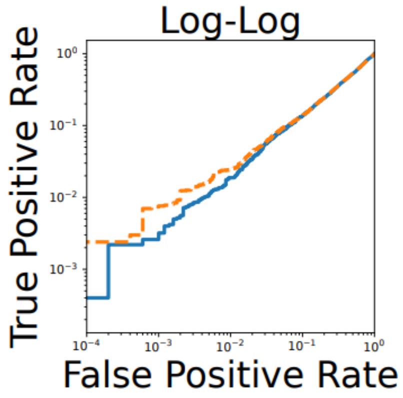
(a) LIRA, CelebA



(b) Logan, CelebA



(c) LIRA, MIMIC-III



(d) Logan, MIMIC-III

Privacy Performance: Area under ROC Curve

TABLE I: Attack AUCROC on CIFAR-10.

	Logan	Ratio	GAN-Leaks
unprotected	0.6435	0.6866	**0.5083**
mixup	**0.5202**	**0.5303**	0.5312
relaxLoss	0.5478	0.5326	0.4197
PAR-GAN	0.6668	0.7398	0.5291

TABLE III: Attack AUCROC on MIMIC-III

	Logan	Ratio	GAN-Leaks
unprotected	0.5264	0.5913	0.5028
mixup	0.5269	0.5283	-
relaxLoss	0.5296	0.5175	-
PAR-GAN	0.5350	0.5015	-

TABLE II: Attack AUCROC on CelebA

	Logan	Ratio	GAN-Leaks
unprotected	0.8346	0.8637	0.5317
mixup	**0.5788**	0.6615	-
relaxLoss	0.7703	**0.5857**	-
PAR-GAN	0.6571	0.7781	-

Some GAN-leaks results are omitted due to poor performance

Utility Performance on Images

TABLE IV: Downstream classification accuracy and FID on Images datasets.

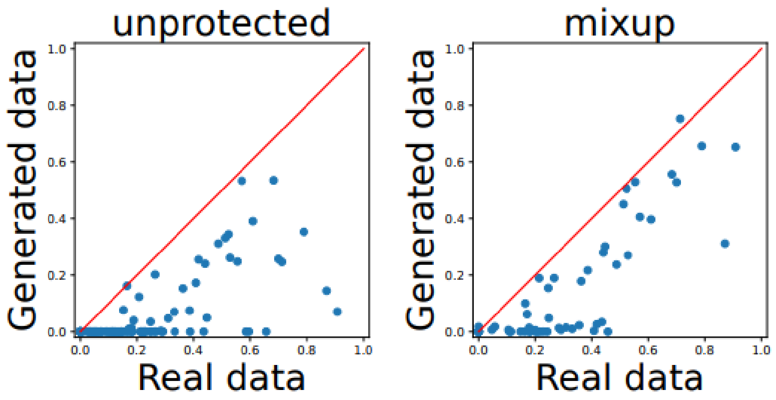
(a) CIFAR-10

Protection	Acc ↑	FID ↓
unprotected	0.490	150.944
mixup	0.421	159.098
relaxLoss	0.385	102.955
PAR-GAN	0.404	199.053

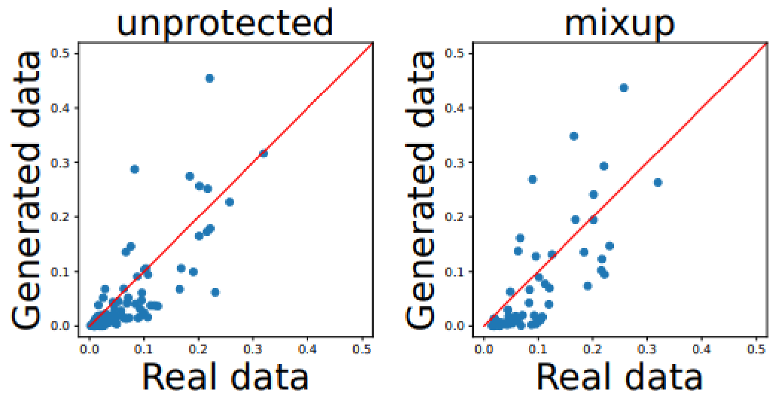
(b) CelebA-Gender

Protection	Acc ↑	FID ↓
unprotected	0.912	111.980
mixup	0.915	104.376
relaxLoss	0.836	97.746
PAR-GAN	0.876	157.724

Utility Performance on Tables



(a) DWpre F1-score of the logistic regression trained on real and generated data



(b) DWP, $\Pr(x_i = 1)$ for each valid i

It can be observed that mixup has a similar utility performance with the unprotected case.

Adaptive Attack

TABLE V: Adaptive attack AUCs against *mixup* on CIFAR-10. The original LIRA against non-protected target GAN has an AUC of 0.6866

query \ ref. models	mixup trained	naturally trained
mixed query	0.5264	0.6084
single query	0.5426	0.5303

TABLE VI: Adaptive attack AUCs against *mixup* on CelebA. The original LIRA against non-protected target GAN has an AUC of 0.8637.

query \ ref. models	mixup trained	naturally trained
mixed query	0.5975	0.7283
single query	0.6701	0.6615

If the attacker knows mixup:
Mixup trained reference model

If the attacker knows the co-membership information:
Mixed query

The strongest one:
naturally trained reference models + mixed samples for co-membership query.

Mixup does provide a significant privacy gain in these cases.

Takeaways

Mixup training can reduce the likelihood ratio for target members.

Mixup training can lower the upper bound of the MIA attacker's AUC.

References

P. Kairouz, S. Oh, and P. Viswanath, “The composition theorem for differential privacy,” in International conference on machine learning. PMLR, 2015, pp. 1376–1385

N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, “Membership inference attacks from first principles,” in 2022 IEEE Symposium on Security and Privacy (SP). IEEE Computer Society, 2022, pp. 1519–1519.

H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” in International Conference on Learning Representations, 2018.

D. Chen, N. Yu, and M. Fritz, “Relaxloss: Defending membership inference attacks without losing utility,” in International Conference on Learning Representations, 2021

J. Chen, W. H. Wang, H. Gao, and X. Shi, “Par-gan: Improving the generalization of generative adversarial networks against membership inference attacks,” in Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 127–137.

D. Chen, N. Yu, Y. Zhang, and M. Fritz, “Gan-leaks: A taxonomy of membership inference attacks against generative models,” in Proceedings of the 2020 ACM SIGSAC conference on computer and communications security, 2020, pp. 343–362.

THANK YOU

SJTU

饮 水 思 源 · 爱 国 荣 校